心理科学进展 2023, Vol. 31, No. 6, 1002–1019
Advances in Psychological Science

https://doi.org/10.3724/SP.J.1042.2023.01002

·研究前沿(Regular Articles)·

语义在人脑中的分布式表征：来自自然语言处理技术的证据*

蒋嘉浩[1] 赵国钰[2] 马英博[1] 丁国盛[3] 刘兰芳[2,4]

([1]北京师范大学心理学部, 北京 100875) ([2]北京师范大学文理学院心理学系, 珠海 519087)
([3]北京师范大学认知神经科学与学习国家重点实验室和IDG/麦戈文脑科学研究院, 北京 100875)
([4]北京师范大学认知神经科学与学习国家重点实验室认知神经工效研究中心, 珠海 519087)

摘 要 人脑如何表征语义信息一直以来是认知神经科学的核心问题。传统研究主要通过人为操纵刺激属性或任务要求等实验方法来定位语义表征脑区, 这类方法虽然取得了诸多成果, 但是依然存在难以详细量化语义信息和语境效应等问题。基于语义的分布式假设, 自然语言处理(NLP)技术将离散的、难以客观量化的语义信息转变为统一的、可计算的向量形式, 极大提高了语义信息的刻画精度, 提供了有效量化语境和句法等信息的工具。运用NLP技术提取刺激语义信息, 并通过表征相似性分析或线性回归建立语义向量与脑活动模式的映射关系, 研究者发现表征语义信息的神经结构广泛分布在颞叶、额叶和枕叶等多个脑区。未来研究可引入知识图谱和多模态融合模型等更复杂的语义表示方法, 将语言模型用于评估特殊人群语言能力, 或利用认知神经科学实验来提高深度语言模型的可解释性。

关键词 语义表征, 大脑, 自然语言处理, 语言模型
分类号 B842

1 前言

语言作为一种抽象符号, 是人类进行意义表达和信息交流的最重要的工具。基于有限数量语言单位的组合, 人们可以理解和表达无穷多的信息, 包括但不限于知识、信念、意图、情感等。揭示人脑如何存储、通达与提取语义一直是认知神经科学的核心问题之一。为了探究语义表征和加工的神经基础, 研究者通常采用的思路是操纵刺激属性或任务要求, 对比不同条件下脑活动模式的异同。例如, 在词汇判断任务中对比真词与假词激活脑区的差异(Pulvermüller, 2013); 或对于相同语言刺激, 对比语义与语音判断任务的脑活动差异(Poldrack et al., 1999)。基于严格实验控制和条件间对比的研究范式取得了一系列重要成果, 然而在探究语义的脑表征与加工问题上存在以下局限。

第一, 对语义特征的刻画依赖人工评定, 且颗粒度较粗。日常生活中交流情境复杂多变, 但人们只需掌握少量的词语即可满足言语交流需求, 例如在汉语中590个字就已经覆盖了80%的日常用字(中华人民共和国教育部, 2013)。有限的文字能够组合成无限的意思, 其原因在于人们对每一词汇都构建了丰富的心理表征, 不同词汇在多个维度上存在微妙差异。基于心理学实验或语言学分类方法, 当前研究对语义关系的度量大多停留在粗颗粒度层面, 例如区分名词与动词, 生命类与非生命类词等。为了细化对语义的表示, 最近有研究者从心理维度对词语概念进行度量, 例如采用时间、空间、数量、唤醒度等12个维度来刻画抽象概念词(X. Wang et al., 2018); 或是采用包括感觉、运动、时间、空间、社会认知等成分在

收稿日期：2022-11-06
* 国家自然科学基金青年科学基金项目(31900802)资助。
通讯作者：刘兰芳, E-mail: liulanfang21@bnu.edu.cn

内的 65 个体验维度来表示概念(Binder et al., 2016)。基于心理维度的语义表示方法能刻画概念本身以及概念间的关系，可解释性较高，但仍具有一定的局限性。例如，维度的选取由研究者主观确定，维度选取的合理性和完整性有待检验。此外，对词义的量化主要通过被试主观判定获得，结果受被试个体知识与经验的影响较大。最后，被试评定法耗时费力，难以推广至所有的词汇，难以全面覆盖不同语境下词语的多个含义，并且不同研究者之间选取的词表与维度有所不同，增加了研究结果间的比较与整合难度。

第二，语境效应难以量化。世界各地的语言系统里，大部分字或词都可指代多种含义，例如在英语中 80%以上的单词都存在一词多义现象(Rodd et al., 2002)。在真实情境下，个体所激活的语言符号含义很大程度上取决于语境，换言之，对语言符号意义的表征和提取是动态的、依赖语境的(Yee & Thompson-Schill, 2016)，例如在夏天和冬天提到“空调”时会倾向于联想到相反的功能。然而，由于语境本身的复杂性，很难通过实验设计手段对语境效应进行客观度量。因此，当前大多数研究使用孤立呈现的语言刺激、打散句法或语义的句子等高度控制的材料，但它们与日常生活中的语言使用相比仍有一段距离。要回答关于人脑如何表征与加工语境，以及语义表征如何受语境信息的动态影响等问题仍面临着较大的挑战。

第三，篇章(discourse)主题信息难以量化。篇章(例如新闻报道、故事)由词和句子以复杂的关系连接而成，不同部分间存在语义关联，能表达完整连贯的含义(主题)。为了探究对篇章语义信息的加工和表征，心理学研究者通常将完整篇章与同一篇章在不同水平(词、句子或段落)打乱后的材料进行对比(Hasson et al., 2008; Lerner et al., 2011; Simony et al., 2016)。然而，打乱后的材料在节点处的复杂度与难度更大(可能引起更强的脑激活)，人们会倾向于尝试重新组织与整合打乱的材料以使其语义连贯，因此条件间相减的方式可能无法准确检测到特异于篇章的语义加工。此外，该实验方法难以度量篇章内不同部分的语义结构关系以及不同篇章之间的语义距离。

鉴于心理学传统实验方法的局限性，近年来越来越多的心理学研究者引入人工智能领域的自然语言处理(natural language processing, NLP)技术，特别是基于人工神经网络和深度学习的语言模型，以度量实验刺激的语义及语义关系。将 NLP 模型与脑成像实验数据相结合，正在成为神经语言学领域的重要趋势。近期有部分国内外研究者对计算语言学方法在认知语言学和脑科学中的应用进行了总结和展望。例如，王少楠等(2022b)总结了新兴计算语言学方法在语言信息的单元和维度、不同类型语言信息的脑网络定位、语言信息加工的时间进程和控制以及语言信息的神经编码形式与计算机制等问题上的应用，文章所探讨的语言信息包括了语音、语义、句法结构等多方面内容。在另一篇文章中(王少楠 等, 2022a)，作者从宏观角度系统地讨论了认知语言学与计算语言学各自的研究问题、研究方法和局限性，并就这两大领域如何融合提出了深刻见解。还有研究者将现代分布式语义计算模型与认知心理学中的两类传统语义模型(基于特征的语义模型和基于联结网络的语义模型)在知识表征、学习机制和语义解歧等方面进行了深入对比，并探讨了现代语义计算模型与两类传统模型的结合途径(Kumar, 2021)。

上述研究在宏观角度概括了计算语言学方法在语言认知中的广泛应用，但未就具体问题进行系统总结和详细论述。本综述拟聚焦语言认知和脑科学领域的核心问题之一——人脑对语义信息的表征，对 NLP 模型在该问题上的应用进行总结与展望。本综述将首先介绍 NLP 模型表征语义的原理与技术，并介绍语言模型与脑成像数据进行结合的两类方法；在此基础上，系统阐述 NLP 技术在人脑语义表征研究中的应用，包括单词语义、句子(及语境)语义和篇章语义，并与传统心理学方法度量语义的局限之处进行对比；最后，探讨应用 NLP 语言模型探究人脑语义表征的潜在陷阱、挑战和未来发展方向。

2　NLP 语义表示的算法原理及进展

如何让计算机从文本中自动捕获语义是计算语言学领域的核心问题之一。早期研究者提出了基于逻辑规则的方法对自然语言进行建模(Chomsky, 1957; Hobbs, 1977)，希望计算机像人一样根据句法、词语顺序和搭配等规则理解词语的含义。尽管该方法的精度较高，但它高度依赖人工编制的语言学文法，不适合处理大规模真实文本(尤其是在词语新用法、新含义越来越多的互联网时代)，

且不同语言之间的规则不尽相同。后来，由于规则表示存在许多问题，统计学派基于“上下文相似的词语，其语义也相似”的分布式语义假设(Harris, 1954)，提出了语义的向量空间模型(Salton et al., 1975)，它成为了 NLP 领域近十余年来的主流指导思想，即分布式表示(distributed representation)。这一思想是把词语这一离散符号(局部表示，local representation)映射到一个稠密的向量空间中，从而使用一个相对低维的向量(例如 300 维)代替稀疏且高达几十万维的独热向量(Bengio et al., 2003)。例如关于颜色的局部表示为“红、橙、黄、灰、中国红……” ([1, 0, 0, 0, 0], [0, 1, 0, 0, 0], [0, 0, 1, 0, 0], [0, 0, 0, 1, 0], [0, 0, 0, 0, 1])，而用分布式表示则可将所有颜色统一到 RGB 三维向量上(例如灰色可表示为[125, 125, 125])，大大减少了向量维度。在分布式表示中，语义信息隐含在词向量的各个维度上，词语间的语义关系主要由它们在空间中的位置关系反映：两个词向量越接近，语义相似性越高。

在语义空间的构建与词向量的获得方面，当前主要有两类思路。一类是基于统计的语义表示方法，该方法主要基于语料库对“词−词”或“词−文档”等的共现关系进行统计，算法包括潜语义分析(latent semantic analysis, LSA, Deerwester et al., 1990; Dumais, 2004)、非负矩阵分解(non-negative matrix factorization, Lee & Seung, 1999)、基于马尔可夫假设的 N-gram (Brown et al., 1992)等。以 LSA 为例，该方法通过统计文本语料建立“词−文档”共现矩阵 $A_{w\times d}$ (其中 w 是词数，d 是文档数)，然后对共现矩阵进行奇异值分解 $A_{w\times d}=U_{w\times r}\Sigma_{r\times r}V_{r\times d}^{T}$ 构建潜语义空间并实现降维(公式中 r 即为潜语义空间维数)。矩阵 U 中每一行为词语的潜语义表示(即词向量)，矩阵 V^T 中的每一列为文档的潜语义表示，矩阵 Σ 中的奇异值反映了每一潜语义的重要程度。如此一来，词和文档的信息得到浓缩，映射到了统一的潜语义空间中，既可以用于词语的语义表示，也可以用于表示篇章和文档的语义。基于统计的语义表示方法能有效聚类语义相近的词或文档，在语义相似性分析、词(或文档)聚类、信息提取等任务上取得了良好的成绩(Jelodar et al., 2019; Xu et al., 2008)。但该方法也具有明显的局限性，例如词(或文档)向量的分布不一定符合概率模型假设所要求的正态分布；矩阵分解的计算复杂度高，并且当加入新的文档时，需重新训练来更新模型；未能充分考虑句子中词语的先后顺序信息；不能解决一词多义现象等。

与基于统计的方法不同，另一类基于预测的语义建模方法使用神经网络学习语义表示，通过计算预测值与真实值的差异来调整模型参数(关于语义建模方法的其他分类标准，请参阅 Kumar, 2021)。人工神经网络(artificial neural network, ANN, 下文简称神经网络)是通过模拟人脑神经系统对复杂信息处理机制而构建的一种数学模型(Mcculloch & Pitts, 1943)。神经网络由神经元(节点)互相连接(边)而构成，按先后顺序主要分为输入层、隐藏层和输出层。输入层主要进行信号接收与激活(例如提取词语对应的词向量，类比于外界刺激引起初级感觉区的电生理活动)；隐藏层是神经网络的核心，主要进行信号的加工、整合和抽象化等复杂过程(类比于大脑中间神经元、联合皮层和高级决策皮层等)；输出层在接收隐藏层加工后的信号后，根据任务需求进行最后一步的反应输出(例如对词语进行情绪分类等，类比于大脑发音皮层、运动皮层)。与大脑神经元动作电位的特性相似，人工神经网络隐藏层中的神经元接收上游多个神经元信号后(类比大脑神经元树突)，按照不同的权重进行加权求和(类比胞体)，随后根据汇总后的信号是否高于激活阈限来决定是否向下游传出信号以及信号的强度(一般经过 sigmoid、ReLU 等非线性激活函数完成)，后续隐藏层的工作过程以此类推。值得注意的是，隐藏层中每个神经元与上游各个神经元之间的信息权重是不同的，这些参数由神经网络输出值与真实值的误差通过反向传播算法不断调整。通过多次训练不断缩小预测值与真实值的差距，神经网络建立起原始输入信号与目标输出间的映射关系，最终的学习结果体现在各个神经元的参数上。

在词语的向量表示问题上，神经网络通常使用大规模语料来训练网络权重，输入句子材料以学习词语和上下文语境的关系。以经典的 Word2Vec 中的连续词袋(continuous bag-of-word, CBOW)模型为例(Mikolov et al., 2013a)，该模型基于分布式假设而设计(上下文相似的词语意思也相似)，给定前后共 k 个上下文语境词，预测中间的目标词。输入层为词的独热编码向量，通过输入层与隐藏层的权重矩阵提取词语的词向量，随后将该向量

与隐藏层输出层之间的权重矩阵进行点乘并使用 softmax 函数进行归一化, 得到词表中各个词出现的概率, 选取概率最高的词语作为预测结果(见图 1)。通过计算预测词与真实词的词向量差异并由反向传播进行参数调整, 输入层和隐藏层之间的权重(即词向量)得以不断更新。此外, Word2Vec 也可以使用跳字模型(skip-gram)进行训练, 即给出一个目标词, 预测其上下文(向前、向后共 k 个词)。Word2Vec 模型获得的词向量与分布式假设吻合较好, 对词向量进行聚类的结果合理, 且能较好地反映语义相似度(Mikolov et al., 2013a; Mikolov et al., 2013b)。例如, 计算向量 V(t) = V(国王) − V(男人) + V(女人), 得到的 V(t)会与 V(女王)等相关词语的词向量余弦相似度最高。

Word2Vec 模型提出以后, NLP 领域掀起了词向量计算与优化表示的热潮, 后续研究者设计了一系列架构更复杂的神经网络语言模型, 它们在计算词向量时考虑了上下文语境的信息, 更符合人脑整合语境的认知模式。新开发的神经网络模型还可以对句子和篇章语义进行建模, 代表性模型包括：可捕获句子的结构信息的递归神经网络(recursive neural network, RecNN, Socher et al., 2013); 循环神经网络(recurrent neural network, RNN, Elman, 1990; Mikolov et al., 2010)及其优化版本长短时记忆网络(long short-term memory, LSTM, Hochreiter & Schmidhuber, 1997; Sundermeyer et al., 2012), 把句子看作一个有顺序的时间序列, 将上(下)文信息整合到当前词语的向量表示中(Graves et al., 2013); 卷积神经网络(convolutional neural network), 提取多层次的语义信息并具备更高效

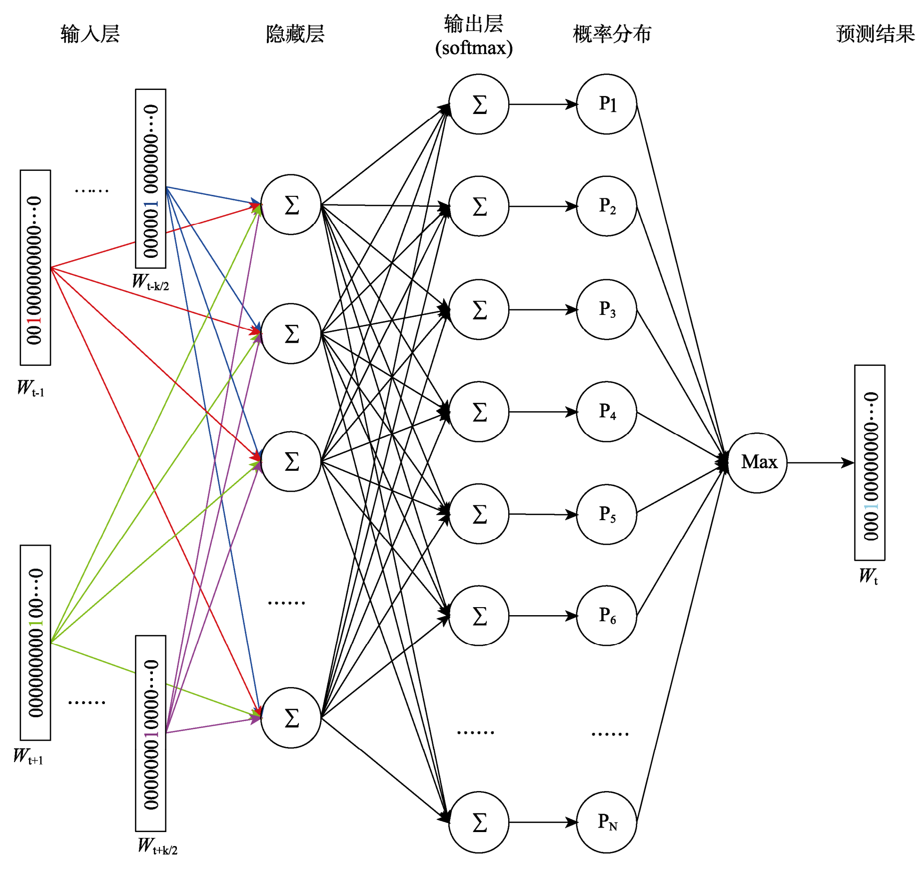


图 1　CBOW 训练示意图

注：对于要要预测的目标词 Wt, 选取向前、向后共 k 个上下文词语作为语境(一般情况下上下文窗口长度相等), 经过输入层和隐藏层的权重矩阵提取它们的词向量进行求和, 随后将新生成的词向量与隐藏层−输出层的权重矩阵进行点乘, 再经过 softmax 计算得到词表(大小为 N)中各个词出现的概率, 选取概率最大的词语作为预测结果。skip-gram 模型略有不同, 输入目标词 Wt, 预测其上下文。

的并行运算能力(Yin et al., 2017; Zhang & Wallace, 2017)。除了词语，基于神经网络的算法也可以对段落或篇章的语义进行表示，例如 Doc2Vec 在 Word2Vec 模型的基础上加入一个段内共享、段间独立的段落向量进行训练，从而获得段落的向量化语义表示(Quoc & Mikolov, 2014)。其他思路还有层次化特征提取等，例如首先计算段落内每句话的语义表示得到句向量，再以句向量为单位输入模型得到段落向量。

后来谷歌公司提出了 Transformer 架构(Vaswani et al., 2017)，解决了 RNN 及其变体的长距离依赖和串行训练低效等局限，成为了近年来 NLP 新模型的主流网络骨干。Transformer 架构由编码器和解码器组成，每个编码器和解码器中包含了自注意力层(multi-head self-attention)和全连接层，其中自注意力层通过对目标词与上下文词语的相似性进行计算与加权求和来整合语境信息，随后经过全连接层提取信息的特征。Transformer 架构中的自注意力机制代替了 RNN 结构中的串行记忆单元，使得计算可以高速并行化，并且 Transformer 架构通过多个编码器和解码器的堆叠提升了对文本特征的提取与抽象效果。基于 Transformer 架构的代表性语言模型包括 BERT (Bidirectional Encoder Representation from Transformers, Devlin et al., 2018)与 GPT (Generative Pre-trained Transformer, Brown et al., 2020; Radford et al., 2019)等，它们在许多自然语言处理任务上的表现都取得了较大的提升。基于深度神经网络的语义建模方法的参数庞大(例如 BERT-large 模型中有 3 亿参数需要训练，GPT-3 的参数量则高达 1750 亿)，对语料数据量、计算机性能等要求较高。因此预训练成为了目前大规模语言模型的主流使用方式，将模型在某个语言任务上进行大量训练(例如完形填空)以得到模型参数，各组研究者以这一套模型参数为基础开展下游任务。预训练模型降低了研究团队训练模型的技术与时间成本，并提升了语言认知研究的可比性与可重复性。

相较于传统基于统计的语义表示方法，神经网络模型能捕获更丰富的文本特征，通用性更强，在完形填空、情感分析、构建文摘、翻译等多种复杂语言任务中具有更优秀的表现(Sutskever et al., 2014; Wu & Dredze, 2019)。此外，大规模预训练模型(例如 BERT)将学习到的多种语言信息都蕴藏在其参数中，研究者可根据自身需要对预训练模型进行微调，从而以较低的资源消耗获得针对专门任务的更优模型表现。随着计算机算力的不断提升，以上优势与表现使得神经网络模型逐步取代传统基于统计的文本表示方法，成为 NLP 领域的核心技术之一。关于 NLP 中的文本表示方法更详细的介绍请参阅赵京胜等(2022)。

3 NLP 语言模型在人脑语义表征研究中的应用

3.1 NLP 语言模型与脑成像数据的结合方法

NLP 语言模型提供了客观度量与计算文本语义的有效工具。利用该工具，神经语言学研究者可以进一步分析语义信息在多大程度上解释了脑活动模式的变化，从而推论出哪些脑区参与了语义信息的表征与加工。值得注意的是，NLP 语言模型得出的词向量与脑活动数据来自不同的模型与模态，各自数据的维度和数值代表的含义截然不同。例如，BERT 输出层的向量为 768 维(BERT-base)或 1024 维(BERT-large)，每一维的数值含义不明确。脑活动的数据维度则根据选取的脑区大小而有所不同，从一维(voxel 水平)，几百(ROI 水平)，几千(网络水平)，到几万(全脑水平)不等。如何对这两类维度不同的多变量数据进行有效建模是一个具有挑战性的问题，当前有两种常用的方法：表征相似性分析(representation similarity analysis, RSA)与线性回归。

RSA 通过分析语义相似性矩阵和脑活动相似性矩阵的共享结构，建立起两类数据的关联(Kriegeskorte et al., 2008)。进行 RSA 分析时，首先需要分别提取人脑和 NLP 语言模型对于各个刺激(例如单词)的表征，其中脑表征可由给定单词引发的一组体素的活动强度数据表示，NLP 模型表征可由 Word2Vec (或其他模型)对该单词的词向量表示。随后分别计算人脑和语言模型内部对于不同刺激的表征相似性程度(可用相关系数、欧式距离或马氏距离等不同指标度量)，从而构建表征差异矩阵(representation dissimilarity matrix, RDM)。RDM 反映了同一个模型对于不同刺激的表征的差异，通过计算两个 RDM 之间的 Spearman 相似性，得到的相关系数反映人脑和语言模型对于同一组刺激的内部表征相似程度(见图 2)。

线性回归是另一种关联不同类型高维数据的

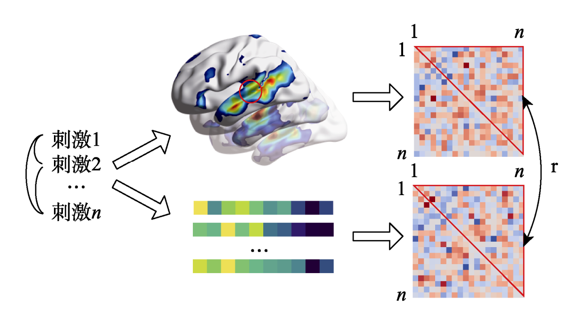


图 2　表征相似性计算示意图。
注：中间上部表示大脑加工不同刺激时的脑活动；中间下部表示用于对比的模型对每个刺激的向量表示。此处的向量既可以是 NLP 模型的词向量，也可以是被试在某些维度上的评分等多种特征。右列是表征不相似性矩阵(RDM)，通过计算脑活动或模型向量在刺激间的两两不相似性得到。计算两个 RDM 上三角的 Spearman 相关系数，即为大脑与模型的表征相似性。

方法，它的基本思想是寻找一组参数去拟合两组数据之间的关系，从而基于刺激特征或模型输出向量“预测”大脑反应(编码)，或基于脑活动模式“预测”被试当前正在加工的内容(解码)。在多种线性回归方法中，岭回归是最常用的一种，它可以解决过拟合与多重共线性等问题。最近有不少研究发现，对于同一语言信息，NLP 模型向量可以通过岭回归与脑活动建立映射关系(王少楠 等, 2022b; Anderson et al., 2021; Caucheteux & King, 2022; Dupre la Tour et al., 2022; Goldstein et al., 2022; Jain & Huth, 2018; Prince et al., 2022; Schrimpf et al., 2021)，若模型和人脑存在相同或相似的表征信息，岭回归预测值与真实值之间将会具有显著相关性。

RSA 和岭回归都可以比较不同模型与脑表征的关系，但它们在原理和功能上有所差异(Bruffaerts et al., 2019)。RSA 度量的是 NLP 模型反应模式与大脑中一组体素(或脑区)反应模式间的相似性程度，而岭回归旨在建立特征(或模型向量)与单个体素(或脑区)活动之间的回归关系。RSA 方法不需要对参数进行拟合，因此计算量小、对数据量要求相对较低。但该方法将所有特征作为一个整体，无法估计单一特征对脑活动的贡献程度。岭回归方法能获取单一特征对脑活动的权重值，进而可根据新刺激的特征预测其激活模式，在使用连续自然刺激的任务中较为常见。但该方法需要估计的自由参数较多，并且往往需要对惩罚系数进行网格搜索，因此计算量较大并且对数据量的要求较高。针对 RSA 和岭回归方法各自的优缺点，Anderson et al. (2016)提出了表征相似性编码方法。该方法基于“相似的刺激会引发相似的脑活动”这一思想，首先计算待预测目标与所有已知目标的特征相似性，随后将相似性指标作为权重对已知目标诱发的脑活动值进行加权平均，从而得到预测目标的脑活动值。该方法利用刺激间的相似性信息进行预测，避免了对模型的参数估计，计算快捷且回归模型中的参数(相似性)具有较强的可解释性，具有较大的应用价值(Anderson et al., 2021; Wang et al., 2020)。值得注意的是，对于 RSA 或岭回归中预测值与真实值的相关系数的解读需要谨慎，显著的相关系数只能说明模型与大脑的表征信息存在相似之处，并不能直接推断二者背后的工作机制是相同的，尤其是相关系数较低的情况下(Kriegeskorte & Douglas, 2018, 2019)。

3.2　典型应用

3.2.1　词水平语义的表征

语言作为思想的载体，其中蕴含的有意义信息由哪些脑区加工、如何加工一直是认知神经科学关注的问题。早期语义表征的研究主要通过比较被试接受不同刺激或进行不同任务时的大脑激活差异，探究词语或概念在哪些脑区进行加工，例如真假词(Pulvermüller et al., 2001)、词语类别(Gonzalez et al., 2006; Pulvermüller et al., 2009)、词性(Pulvermüller et al., 1999; Warburton et al., 1996)、语义任务和语音任务(Poldrack et al., 1999)的对比等。条件对比范式与激活分析取得了不少重要发现，但对语义信息的刻画主要停留在粗颗粒度层面且难以量化。NLP 技术使得研究者能从定量角度对材料的语义信息进行度量，探究语义信息与大脑表征之间的关联。

在早期的工作中，Mitchell 等人(2008)选取名词刺激作为材料，使用它们与 25 个代表性动词的共现频率作为语义向量表示，通过线性回归对大脑加工名词时的活动进行预测。结果发现双侧枕叶、顶叶、额中回等区域都能够区分词语，说明大脑对实体名词的表征一定程度上基于感觉运动特征，其中枕叶的效应可能是因为被试对名词的相关动作场景产生了联想。该研究开创了 NLP 与脑成像技术相结合的先河，为语义脑表征研究提供了条件对比范式以外的新思路。近期研究者开

始将 NLP 方法应用到对自然连续语言材料(例如故事或电影音频)的语义分析中(Huth et al., 2016; Wehbe et al., 2014), 相比于传统的实验室方法(人为编制或挑选少量特定的语言刺激), 这些自然连续材料包含的词汇量更大、类型更丰富, 因此得出的结果可能更能反映真实的人脑语义表征。例如在 Huth 等人(2016)的研究中, 被试收听了长达 2 个小时的故事并同步进行 fMRI 扫描。研究者首先标记每个 TR (repetition time)内出现的刺激, 提取这些刺激对应的词语共现向量作为该 TR 的语义表示, 随后构建岭回归预测模型, 使用语义表示向量预测大脑每个体素的活动。若某个体素的预测相关性经过多重比较校正后依然显著, 说明它的活动蕴含了语义信息, 即参与了语义表征。结果表明, 语义信息在大脑中的表征分布覆盖了内侧前额叶、颞中回、颞顶联合区等多个脑区, 与元分析发现的语义网络(Binder et al., 2009)高度重叠。这些研究成果表明 NLP 对语义的表示能够有效地运用在复杂的自然刺激中, 并进一步支持了语义的分布式表征观点(Kiefer & Pulvermüller, 2012; Nastase et al., 2017), 即多个脑区共同加工、表征语义, 而非集中在某一局部区域内。

此外, NLP 技术对词汇语义的量化功能使研究者能够从更精细的角度考察语义表征, 拓宽了研究空间。例如 Kivisaari 等人(2019)考察了人们对概念的表征与概念特征之间的联系, 在研究中向被试逐一呈现目标概念的 3 个特征词(例如“一种水果”、“被剥开”、“猴子吃它”), 被试需要根据这些特征猜想对应的概念(例如“香蕉”)。研究者使用大脑体素活动模式对特征词或目标词的词向量进行解码, 并比较蕴含不同信息的词向量的解码正确率。结果表明, 尽管被试只看到了 3 个特征词, 但将目标概念的所有特征(包括没有呈现的特征)对应的词向量相加后得到了最高的解码正确率, 显著高于呈现的特征词语和目标概念, 说明人脑利用有限的信息片段构建了目标对象完整的语义表征, 并且激活了其他相关联的概念特征信息。

3.2.2 语境信息的影响以及句水平语义表征

在探究语义在大脑中的表征时, 许多研究将词语或目标刺激单独呈现, 希望获得没有其他信息干扰下的语义表征。然而语义表征是动态的(Yee & Thompson-Schill, 2016), 同一词语在不同的语境中表达的意思和产生的心理感受会有所不同。例如人们看到“女排”一词的心理表征与“中国女排”会有所不同, 后者的“女排”在“中国”语境下可能会激活自豪感、具体的人物形象等额外信息。已有研究表明, 颞叶前部、额顶网络等脑区会整合并更新当前的语义信息(Bonnici et al., 2016; Branzi et al., 2020; Humphreys et al., 2021; Lambon Ralph et al., 2017), 进一步说明了语义表征的动态性。语境独立的实验设计或静态词向量并不能充分地刻画丰富语境下的语义表征, 尤其是面临一词多义现象时。

NLP 技术提供了能够整合语境的多种深度语言模型, 例如 ELMo (Peters et al., 2018)、InferSent (Conneau et al., 2017)、BERT 等, 对于同一个词, 模型输出的语义向量能随着语境的不同而变化。利用该特点, 有研究者比较了孤立词和整合语境信息后的词在人脑中的表征(Gao et al., 2023)。在实验中, 每个试次包含两个先后呈现的英语单词, 被试需要判断它们是否存在语义关联。研究者首先采用 Word2Vec 模型提取语义向量, 该模型对词的语义表示是相对固定的, 不受情境词的影响, 因此被认为反映了词的孤立语义。同时, 对于同一单词, 研究者还采用了 ELMo 模型提取其语义向量, 该模型采用双向循环神经网络结构, 输出的词向量充分整合了语境信息(即前一个词)。通过使用 RSA 比较人脑和语言模型对于同一组刺激的内部表征相似程度, 研究者发现孤立语义的表征主要与缘上回有关, 而语境依赖的语义表征则主要与左侧前额叶、角回和腹侧颞叶有关。

通过运用自注意力机制整合上下文语境信息, NLP 技术还提供了表征句水平语义的指标(例如 InferSent 模型的输出向量或 BERT 模型输出的 CLS 向量)。句水平的向量表示不仅考虑了单个词的语义信息, 还考虑了词与词之间的组合关系。在近期一项研究中, 被试观看一系列由 4~9 个单词构成的句子, 同时进行 fMRI 扫描。研究者首先使用 InferSent 模型提取句子的语义表征, 然后通过岭回归建立句子语义特征与脑活动模式间的预测关系。结果发现, 表征句义的相关脑区分布在包括额下回、额中回、颞上回、颞中回、枕中回在内的广泛区域(Anderson et al., 2021)。在另一项研究中, 被试观看电影的同时进行 fMRI 扫描。研究者将电影切割成多个片段, 并对每一片段进行文字注释(每条注释大约包含 15 个词), 然后采用 NLP

模型将注释转换成语义向量作为电影片段的语义特征，最后基于脑活动数据预测各个片段的文本注释语义特征。研究表明，默认网络、语言网络、枕叶的脑活动模式能较为准确地预测片段语义特征并区分不同的片段(Vodrahalli et al., 2018)。与上述研究结果一致，Acunzo 等人(2022)首先训练一个对话题进行分类的卷积神经网络以使模型向量更好地捕获话题信息，随后提取该模型的输出层向量作为句子的话题向量表示。将话题向量与大脑活动进行表征相似性分析发现，颞叶前部、默认网络等参与了话题水平信息的表征，支持了默认网络具有抽象、整合长时程信息等意义建构功能的观点(Smallwood et al., 2021; Yeshurun et al., 2021)。

3.2.3 分离句法和语义

语言信息能够顺利传达，不仅依赖词语本身的语义信息和语境提供的背景信息，还需要词语之间有恰当的组织结构，即句法。经典的句法研究范式主要采用对比的思路试图分离句法加工成分，例如将名词、形容词等内容词替换成假词的 jabberwocky 句式(Fedorenko et al., 2012; Matchin et al., 2019)、句法违背(Batterink & Neville, 2013; Petersson et al., 2012)、句法适应(Segaert et al., 2012)以及短语组合(Law & Pylkkanen, 2021)等。然而传统的句法加工研究方法存在着一些局限，例如不同任务得到的句法加工脑区分布有不少差异，并且由于语义和句法总是相伴出现，改变句法而不使语义发生变化有一定的难度(Pylkkanen, 2019)，因此句法错乱的句子很大程度上破坏了语义信息，使得传统实验难以分离精细的句法加工过程(Kuperberg, 2007)。

自然语言文本中词语的顺序结构蕴含了丰富语言信息，即使没有显式表示句法关系，具有语境整合能力的 NLP 模型在训练过程中也会习得句法关系，例如“我”、“爱”、“你”会以“我爱你”的顺序出现，而不是“我你爱”。深度语言模型(例如 BERT)在主谓一致性、反身代词回指等多种句法任务上已经接近甚至超越人类表现(Goldberg, 2019; Zhang et al., 2022)，表明其能够较为准确地从文本中获取句法信息。采用实验设计中“减法”的思路，可以使用 NLP 模型分别提取句子中的句法和语义信息，将句法信息从向量中剥离，从而探究加工句法信息的脑区分布(Caucheteux et al., 2021a, 2021b; Wang et al., 2020)。研究结果发现，双侧颞叶和额下回都对句法信息进行了加工，脑区分布情况与先前的实验研究相似(Hagoort & Indefrey, 2014)。最近有研究者使用特征消除(feature elimination)的方式对句法信息进行更精细的分离(例如词性、命名实体、词语依赖、语义角色等)，进而探究被试在倾听故事时所进行的多种句法加工(Zhang et al., 2022)。结果发现，尽管不同句法对应的脑区分布有细微的差异，但分布的区域大致相同，集中在颞上回、颞中回和角回等语义网络区域(Binder et al., 2009)。

NLP 模型可以有效地分离语义和句法信息，并能够在限制较少的自然任务中探究大脑的加工机制，这两大优点预示着 NLP 模型在脑表征研究方向上的潜力(Cichy & Kaiser, 2019; Hamilton & Huth, 2020)。然而，目前使用 NLP 模型探究大脑句法加工的研究数量有限，其中发现的句法加工脑区比传统研究方法覆盖了更广的区域，这一现象究竟是对分布式句法信息加工机制的如实反映，还是源于 NLP 模型与脑成像数据构建映射时存在的误差，仍需将来研究开展进一步分析。

3.2.4 篇章主题信息与篇章语义结构的表征

篇章(段落)理解建立在词和句子的语义分析基础之上，通过识别篇章内部不同部分的语义结构关系、整合上下文信息，最终形成篇章核心主题信息(或情境模型)的表征(Patel et al., 2022)。传统实验方法一般将完整篇章与打乱的材料进行对比(Hasson et al., 2008; Lerner et al., 2011; Simony et al., 2016)，而散乱的材料使得被试的记忆与整合难度更大，因此探测到的差异可能并非完全由特异于篇章语义信息的加工所驱动。此外，该方法未对篇章信息进行量化，难以度量篇章间的语义距离与关系，不适用于不同篇章材料的研究。

近年来已有研究者开始利用 NLP 技术对篇章的语义进行建模表示，考察人脑对连续自然语言刺激(如故事或电影)的加工和表征。近期一项研究结合 fMRI 技术和 LSA 方法，探究以不同模态呈现的复杂叙事信息在人脑中如何表征(Nguyen et al., 2019)。实验中被试在接受 fMRI 扫描的同时，其中一组观看无声影片，另一组收听影片内容对应的语音叙述。在扫描结束后被试用自己的话描述故事内容，研究者通过 LSA 进行语义分析，发现不论观看无声影片还是收听语音叙述，被试描述

内容的语义相似度越高，他们在默认网络与执行控制网络上的神经活动相似度也越高，这一研究结果揭示了默认网络(default mode network, DMN)跨模态表征主题语义信息的功能。另一项研究考察了言语产生和言语理解过程中大脑对主题信息的表征一致性(Patel et al., 2022)，在 fMRI 扫描的同时，被试围绕一系列主题进行口头描述，并收听另一被试讲述的其他主题内容。研究者运用 LSA 计算描述内容两两之间的语义距离，并计算言语理解任务和言语产生任务的脑表征差异矩阵，最后计算语义差异矩阵和脑表征差异矩阵的相似度(RSA 分析)。结果表明，包括额下回、内侧前额叶、颞极、颞中回、角回和楔前叶在内的双侧广泛脑区，其活动模式与言语理解和产出的语义内容存在关联。该研究首次对言语产生过程的篇章水平语义进行分析，揭示了言语产生和言语理解两个过程共享的负责高层级篇章语义信息表征的网络。以上研究通过对篇章水平语义信息进行分析，研究结果进一步支持了默认网络在意义构建中的作用(Margulies et al., 2016; Smallwood et al., 2021)。

对篇章材料还可以从网络拓扑属性方面探究语义结构对大脑加工、学习、记忆等的影响。在文本、视频等自然刺激当中，句子和事件在某一主题内是相互联系的，例如一个故事通常围绕着若干个核心的主旨句或情节进行展开。使用语义相似性作为连边的权重，对篇章构建拓扑网络，可以反映篇章的语义组织结构等信息。有研究者对电影叙事节奏与观众评价之间的联系进行探究(Laurino Dos Santos & Berger, 2022)，使用相邻片段的语义相似性作为衡量情节发展速率的指标，情节发展缓慢时相邻片段的语义相似度较高。研究结果显示，开头节奏缓慢、结尾情节推进稍快的电影得到了更高的评分，表明故事篇章的语义结构会对人们的感受与投入度产生影响。最近另一项脑成像研究考察了篇章语义结构对记忆效果的影响(Lee & Chen, 2022)，研究者对视频片段进行分割，借助 NLP 技术提取各个片段对应文字描述的语义向量，并以片段作为节点、以片段间的语义相似性作为连边权重，构建视频的语义结构拓扑网络。研究结果显示，中心度(centrality, 反映了与其他节点的关联强度)较高的片段产生了更好的记忆效果，并且在情景回忆相关脑区(默认网络)诱发了更强的激活与更高的被试间一致性，表明人脑对于事件的加工与记忆效果与其在语义组织结构中的位置有关。

以上研究结果表明篇章的语义组织结构对人们的主观感受、记忆效果与大脑活动等都产生了影响，但目前使用 NLP 对大脑语义表征的研究大多从刺激编码角度出发，对连续刺激中的语义组织结构和语义关系等关注较少。未来研究可以从自然刺激中的语义结构入手，进一步探究其与大脑加工、学习和记忆效果的关联，例如对于阴谋论和谣言的识别(Miani et al., 2022)、叙事偏好(Cooper & Nisbet, 2016)等的神经基础。

3.2.5 小结

NLP 技术的使用让语言从符号表示转为向量表示，一定程度上克服了词语离散、难量化计算、难统一表示等难点，使得语义的计算和比较成为可能。与此同时，表征相似性分析、线性回归等多变量分析方法为不同模态的数据搭建了桥梁。随着深度语言模型的发展，NLP 模型已能够将上下文语境信息整合进向量表示中，提升了对语言的表示精度，并使得实时刻画语义在不同语境背景下的动态变化成为可能。基于此，研究者使用 NLP 提取的词向量作为语义表示，减少了对于刺激材料或实验任务等的人为控制需求，对语义脑表征的探究不再依赖不同类型刺激或加工任务的对比。此外，NLP 作为计算语言模型具有较高的灵活性，输入不同类型的文本可以得到对应的信息。研究者可以通过比较模型对不同类型文本的向量表示(例如含语境信息的词向量和不含语境信息的词向量)与大脑表征的匹配程度，分析某一脑区表征的信息类型或加工特点(Cichy & Kaiser, 2019)，例如人脑对未来词语的预测机制(Caucheteux et al., 2021b; Goldstein et al., 2022)，先验信念对文本理解的影响(Tikochinski et al., 2021)等。通过将实验设计的对象从大脑活动转移到计算模型上，NLP 技术可用于分离不同成分的信息，并有效降低了被试与实验数量的要求。最后，自然刺激和低限制任务的使用正逐渐成为脑成像研究的趋势(Finn & Bandettini, 2021; Hamilton & Huth, 2020)，然而传统心理学实验方法难以追踪不断输入的词语语义、难以将先前语境信息整合到当前词语中。NLP 技术提供了表征字、词、句、篇章等多层级语义信息的建模方法，在自然语言加工的脑神经基础探究中发挥着日益重要的作用。

运用NLP技术提取刺激的语义特征并与脑活动建立映射关系，近期研究者较为一致地观察到语义表征有关的神经结构广泛分布在额叶、颞叶、枕叶等多个脑区，该结果与基于传统心理学实验方法以及脑损伤病人所揭示的局部脑区表征语义结论并不完全一致。其部分原因可能在于，基于大样本文本库训练得到的语言模型较为充分地捕获了语言符号的多重语义信息，而传统心理学实验中使用的特定任务(例如：语义关联判断)选择性地激活了语言符号某一方面的语义，因而以往仅探测到部分脑区的参与。值得注意的是，有不少理论模型也提出语义记忆的神经表征分布在包括感觉运动区和联合皮层在内的广泛脑区(Bi, 2021; Fernandino et al., 2016a; Fernandino et al., 2016b; Lambon Ralph et al., 2017)。例如，概念表征的“中心−辐射(hub-and-spoke)”理论(Patterson et al., 2007; Lambon Ralph et al., 2017)提出，跨通道的语言及非语言经验构成了概念的核心成分(即 hub)，主要由颞叶前部负责表征与整合，而概念习得过程中出现的初始源头信息(即 spoke，包括视觉、听觉、情绪效价等)则分布在各个通道特异皮层。此外，双重编码理论则将知识表征分为两大类别：基于感知运动(sensorimotor-derived)的系统与基于语言符号(language-derived)的系统，其中支持感知运动编码的知识表征系统主要分布在通道特异的感觉运动皮层以及联合皮层等广泛脑区；支持语言编码的知识表征系统则主要分布在背侧前颞叶(dorsal anterior temporal lobe, dATL)及其延展区域(包括额下回和颞中回等经典语言脑区)。基于NLP技术揭示的广泛语义敏感脑区说明表征语义的向量空间有可能同时捕获了自然语言的抽象、跨通道成分和通道特异成分，然而要建立起这些研究发现与认知理论模型之间的确切关联还面临着众多挑战(关于该问题更深入的讨论请参阅：王少楠等, 2022a; Kumar, 2021)。

4　总结与展望

相比传统心理学实验方法，运用自然语言处理(NLP)技术来刻画语义具有几大优势：(1)能够对词、句子和篇章等多个层级的语义信息进行客观量化和计算，提供了语义的度量指标；(2)能够整合上下文信息，根据语境调整词向量的输出，从而对语境下的语义有更准确的表示；(3)NLP 模型输出的词向量蕴含丰富的信息，通过消融实验或输入不同类型刺激等方式，研究者可以提取或去除某种信息(例如句法信息)，从而在不同的信息角度对大脑语义表征进行考察；(4)词向量的获取快速便捷、受主观因素干扰较少，能大大降低材料评定所需成本。通过表征相似性和线性回归等方法，研究者尝试利用基于语言模型提取的语义信息来解释脑活动的变化，在揭示语义的分布式表征、语境信息对语义表征的影响、句法与语义加工区域的分离以及篇章语义表征等问题上取得了诸多新发现。

然而，在回答语言认知及其脑机制等相关问题时，自然语言处理技术也存在一定的局限性。首先是 NLP 模型的可解释性问题。近年来基于神经网络和深度学习技术的语言模型内部结构越来越复杂和庞大，例如最近的 GTP-3 模型参数量达到了 1750 亿(Brown et al., 2020)，尽管在语言任务上的表现较好，但庞大的参数量和复杂的结构使得模型的可解释性较差：模型输出的词向量反映了语言哪些方面的特征？模型通过哪些关键步骤获得了这些特征？这些问题目前尚无确切答案。目前可以采用模型对比等方式(例如消除或保留语境信息、采用随机向量代替词向量等)探究大脑对某种信息的加工，但低可解释性仍然在一定程度上限制了 NLP 在语言认知研究上的解释效力与应用潜力。其次，模型的数量和类型正迅速增长，不同模型在训练材料、网络架构、参数量以及训练任务等多个方面存在差异，导致输出的词向量不尽相同。在使用词向量与大脑活动建立映射关系时，模型之间编码或解码的表现差异来源变得模糊，即使采用相同的预训练模型来获得相同的模型参数，也面临着模型抽样误差等问题。此外，NLP 模型的构建与人类习得语义的途径不同，其内在计算与加工机制也可能与人脑存在本质差异。人类的语言习得是不断与世界环境进行多模态交互的过程，而目前主流 NLP 模型绝大多数只有文字一个模态，并且难以做到像人类一样基于短短几次反馈就习得新知识或改变原有观念。另一方面，NLP 模型的训练语料越来越多、结构越来越复杂，在逻辑推理、知识迁移等高级语言任务上却仍然表现较差，NLP 是否真正习得语言目前是一个备受关注的问题。因此，借助 NLP 模型能够多大程度解释人脑中的语义表征机制仍需未来

更深入的研究。鉴于以上局限性，在应用语言模型提取刺激特征时，研究者需根据研究问题选择恰当的模型，结合实验设计对模型的有效性进行测试，并谨慎解释实验结果。

值得注意的是，NLP 模型并不总是语义表示的唯一解或最优解。当前心理学的其他语义表示方法在一些情况下也取得了不错的表现，并且具有较强的可解释性，例如特征列举法能够直观地反映概念不同特征在记忆中的凸显度(Cree & McRae, 2003)；特征评定法能获得概念在多个维度(例如感知觉、情绪等)上的属性强弱，也能以分布式表示对概念进行相似性等计算(Binder et al., 2016)；网络模型能够清晰地反映概念之间的层级与关系结构(Solomon et al., 2019; Zhu et al., 2022)。基于纯文本进行训练的 NLP 模型并不一定能完整捕获人类的语义知识以及加工特点(如推理、联想、多模态等)，例如最近对概念语义脑表征的研究发现，相比于 NLP 模型，基于体验属性的特征评分与大脑的表征相似性更高，并且使用偏相关控制共享信息的影响之后，体验属性仍表现出与大脑显著的表征相似性，而 NLP 模型则相关不显著，说明人脑对概念的表征中存在 NLP 模型尚未学习到的多模态信息(Fernandino et al., 2022; Tong et al., 2022)。因此，NLP 模型与传统的心理学语义表示方法并无绝对的优劣之分，它们提供了互补的信息与作用(Kumar, 2021)：在小规模语料中，传统方法虽然颗粒度较粗，但其高解释性有助于对研究理论与假设进行验证；在大规模语料和自然刺激中，虽然 NLP 模型较低的可解释性使得向量维度含义不明确，但其能够便捷地获取语境化的语义表示，并通过模型对比的方式对不同信息内容进行考察。

下一步，研究者还可从以下几个方面深入拓展 NLP 技术在神经语言学中的应用：

(1)引入基于图模型的语义表示方法。除了基于分布式假设的文本表示方法，图模型也是 NLP 领域中较为成熟的表示文本关系的技术(例如知识图谱)。在图模型中，网络的节点代表语言要素(词、概念、实体、句子、篇章等)，网络的边代表语言要素间的关系。以知识图谱为例，图模型的建构充分利用了语言要素的属性关系、语言学先验知识和世界知识等信息，与神经网络模型相比具有更高的可解释性，语义关系明确，易于进行常识推理任务。但图模型用于表示语义的数据结构较为复杂，难以直接使用图模型的语义表示对脑活动数据进行直接建模，研究者可采用间接的方式，从图模型中提取语义关系或距离信息，随后使用 RSA 等方法考察大脑对语义关系的加工。以 WordNet 为例，该数据库根据单词间的语义关系(例如从属关系)，将单词按照树状结构进行组织。WordNet 中两个词之间的语义距离可通过连通这两个词所需的最短路径来度量(Carota et al., 2021; Fernandino et al., 2022; Wurm & Caramazza, 2019)，例如，从“猫(cat)”的节点到达“鼠(mouse)”的节点需要经过以下路径：猫—猫科动物—食肉动物—胎盘哺乳动物—啮齿类—鼠，因此这两个词的关系距离为 5。

(2)应用多模态融合的深度语言模型。在自然交流情境下，人们对信息的加工与理解常常融合了声音、图像、文本等多个模态，并且加工单个概念时往往也会提取多个模态的信息(Bi, 2021)，然而传统的实验方法和基于纯文本的 NLP 模型难以融合与量化多模态信息，尚不能全面描述人脑对于概念的表征内容(Dubova, 2022; McClelland et al., 2020)。人工智能领域已经开发了多模态融合的深度语义表示方法(Lahat et al., 2015; Wang, Zhang, Lin et al., 2018; Wang, Zhang, Zong, 2018; Zhu et al., 2022)。运用多模态语言模型，可进一步深入探究大脑对不同模态信息的加工机制，例如基于语言和基于体验的两类信息(Bi, 2021; Paivio, 1991)在大脑中的表征分布与方式、角色地位以及整合方式与程度等。

(3)运用语言模型评估特殊人群的语言能力。例如，对正常人和失语症(或自闭症、精神分裂症等)患者的语言产出进行文本分析，获得其语义类别、语义模糊性、词频分布和语义结构等多方面特征(Day et al., 2021; Nevler et al., 2020)。基于这些特征建立分类或预测模型，有助于提高语言能力与疾病评估的准确性或受测者的接受程度(de Boer et al., 2018; Fraser et al., 2016)，并降低评估所需的时间与人力成本。

(4)利用脑活动数据增强对深度语言模型的理解或改进模型。现今的深度语言模型能完成各种各样的语言任务，但人们对模型内部的实现机制依然缺乏清晰的认识。人脑是世界上唯一能真正理解自然语言的加工系统，理解深度模型的一个

思路便是将其与人脑进行对比，当前已有部分研究开始基于深度模型的“类脑”情况来推测模型内部的运行机理或解释不同模型存在差异的原因。例如在一项研究中，研究者拟探究不同语言模型以及同一模型内部不同隐藏层对语境信息的整合能力(Toneva & Wehbe, 2019)。研究者使用 fMRI 采集了被试阅读故事(每个词单独呈现在屏幕上)时的脑活动，同时提取了不同 NLP 模型的每一隐藏层对故事中每个词的向量表示，通过岭回归和分类任务计算模型输出词向量对多个重要语言脑区活动的预测程度。结果表明，当用于计算词向量时纳入的语境较短时(少于 10 个词)，BERT 和 Transformer T-XL 模型的中间层对脑活动的预测效果优于较浅的输入层，反映了隐藏层的语境整合能力。当纳入的语境信息超过 10 个词时，BERT 对脑活动的预测效果随着语境词数量的增加而下降，而 Transformer T-XL 的预测效果则仍然保持缓慢升趋势。研究者推测，对脑活动预测效果最佳时对应的语境长度可能反映了模型(或隐藏层)整合情境信息的能力，结果显示 Transformer T-XL 比 BERT 更擅长整合长距离语境信息，而这正是 Transformer T-XL 当初的设计初衷之一。类似的工作还发现 NLP 的语言任务能力和对大脑活动的预测能力存在显著正相关(Caucheteux & King, 2022; Schrimpf et al., 2021)。更进一步，还有研究者对模型进行微调，发现提高模型对脑活动预测能力的同时(使模型更“类脑”)显著改善了模型在多个语言任务上的表现 (Schwartz et al., 2019; Toneva & Wehbe, 2019)。

以上研究表明，通过与人脑的认知和神经加工过程作对比来理解甚至改进深度语言模型这一方向具有很大潜力。但由于人类思维的隐蔽性和当前脑成像技术在时间和空间分辨率上的局限性以及低信噪比等问题，进行“类脑”分析或对 NLP 模型内部的认知机制进行探究时，仍需利用严格的实验控制和先验知识对结果进行约束，或配合其他模型解释方法共同做出推理(Sun et al., 2021)。

参考文献

王少楠, 丁鼐, 林楠, 张家俊, 宗成庆. (2022a). 语言认知与语言计算——人与机器的语言理解. *中国科学：信息科学*, *52*(10), 1748−1774. https://doi.org/10.1360/SSI-2021-0100

王少楠, 张家俊, 宗成庆. (2022b). 基于语言计算方法的语言认知实验综述. *中文信息学报*, *36*(4), 1−11.

赵京胜, 宋梦雪, 高祥, 朱巧明. (2022). 自然语言处理中的文本表示研究. *软件学报*, *33*(1), 102−128. https://doi.org/10.13328/j.cnki.jos.006304

Acunzo, D. J., Low, D. M., & Fairhall, S. L. (2022). Deep neural networks reveal topic-level representations of sentences in medial prefrontal cortex, lateral anterior temporal lobe, precuneus, and angular gyrus. *NeuroImage*, *251*, 119005. https://doi.org/10.1016/j.neuroimage.2022.119005

Anderson, A. J., Kiela, D., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., ... Lalor, E. C. (2021). Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *Journal of Neuroscience*, *41*(18), 4100−4119. https://doi.org/10.1523/JNEUROSCI.1152-20.2021

Anderson, A. J., Zinszer, B. D., & Raizada, R. D. S. (2016). Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage*, *128*, 44−53. https://doi.org/10.1016/j.neuroimage.2015.12.035

Batterink, L., & Neville, H. J. (2013). The human brain processes syntax in the absence of conscious awareness. *Journal of Neuroscience*, *33*(19), 8528−8533. https://doi.org/10.1523/jneurosci.0618-13.2013

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*(6), 1137−1155. https://doi.org/10.1162/153244303322533223

Bi, Y. (2021). Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, *25*(10), 883−895. https://doi.org/10.1016/j.tics.2021.07.006

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, *33*(3-4), 130−174. https://doi.org/10.1080/02643294.2016.1147426

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767−2796. https://doi.org/10.1093/cercor/bhp055

Bonnici, H. M., Richter, F. R., Yazar, Y., & Simons, J. S. (2016). Multimodal feature integration in the angular gyrus during episodic and semantic retrieval. *Journal of Neuroscience*, *36*(20), 5462−5471. https://doi.org/10.1523/jneurosci.4310-15.2016

Branzi, F. M., Humphreys, G. F., Hoffman, P., & Lambon Ralph, M. A. (2020). Revealing the neural networks that extract conceptual gestalts from continuously evolving or changing semantic contexts. *NeuroImage*, *220*, 116802, Article

116802. https://doi.org/10.1016/j.neuroimage.2020.116802

Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*(4), 467−480. https://aclanthology.org/J92-4003

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Askell, A. (2020, December). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877−1901. https://dl.acm.org/doi/10.5555/3495724.3495883

Bruffaerts, R., de Deyne, S., Meersmans, K., Liuzzi, A. G., Storms, G., & Vandenberghe, R. (2019). Redefining the resolution of semantic knowledge in the brain: Advances made by the introduction of models of semantics in neuroimaging. *Neuroscience & Biobehavioral Reviews*, *103*, 3−13. https://doi.org/10.1016/j.neubiorev.2019.05.015

Carota, F., Nili, H., Pulvermuller, F., & Kriegeskorte, N. (2021). Distinct fronto-temporal substrates of distributional and taxonomic similarity among words: Evidence from RSA of BOLD signals. *NeuroImage*, *224*, 117408, Article 117408. https://doi.org/10.1016/j.neuroimage.2020.117408

Caucheteux, C., Gramfort, A., & King, J.-R. (2021a, July). Disentangling syntax and semantics in the brain with deep networks. *Proceedings of the 38th International Conference on Machine Learning*, *139*, 1336−1348. https://proceedings.mlr.press/v139/caucheteux21a.html

Caucheteux, C., Gramfort, A., & King, J.-R. (2021b). Long-range and hierarchical language predictions in brains and algorithms. *arXiv*. https://doi.org/10.48550/arXiv.2111.14232

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134. https://doi.org/10.1038/s42003-022-03036-1

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton. https://doi.org/10.1515/9783112316009

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*(4), 305−317. https://doi.org/10.1016/j.tics.2019.01.009

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017, September). Supervised learning of universal sentence representations from natural language inference data. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 670−680, Copenhagen, Denmark. https://doi.org/10.18653/v1/D17-1070

Cooper, K. E., & Nisbet, E. C. (2016). Green narratives: How affective responses to media messages influence risk perceptions and policy preferences about environmental hazards. *Science Communication*, *38*(5), 626−654. https://doi.org/10.1177/1075547016666843

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, *132*(2), 163−201. https://doi.org/10.1037/0096-3445.132.2.163

Day, M., Dey, R. K., Baucum, M., Paek, E. J., Park, H., & Khojandi, A. (2021, November). Predicting severity in people with aphasia: A natural language processing and machine learning approach. *Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, *2021*, 2299−2302, Mexico. https://doi.org/10.1109/embc46164.2021.9630694

de Boer, J. N., Voppel, A. E., Begemann, M. J. H., Schnack, H. G., Wijnen, F., & Sommer, I. E. C. (2018). Clinical use of semantic space models in psychiatry and neurology: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *93*, 85−92. https://doi.org/10.1016/j.neubiorev.2018.06.008

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391−407. https://doi.org/10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://doi.org/10.48550/arXiv.1810.04805

Dubova, M. (2022). Building human-like communicative intelligence: A grounded perspective. *Cognitive Systems Research*, *72*, 63−79. https://doi.org/10.1016/j.cogsys.2021.12.002

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, *38*(1), 189−230. https://doi.org/10.1002/aris.1440380105

Dupre la Tour, T., Eickenberg, M., Nunez-Elizalde, A. O., & Gallant, J. L. (2022). Feature-space selection with banded ridge regression. *NeuroImage*, *264*, 119728. https://doi.org/10.1016/j.neuroimage.2022.119728

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179−211. https://doi.org/10.1207/s15516709cog1402_1

Fedorenko, E., Nieto-Castanon, A., & Kanwisher, N. (2012). Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia*, *50*(4), 499−513. https://doi.org/10.1016/j.neuropsychologia.2011.09.014

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W. L., ... Seidenberg, M. S. (2016a). Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral Cortex*, *26*(5), 2018−2034. https://doi.org/10.1093/cercor/bhv020

Fernandino, L., Humphries, C. J., Conant, L. L., Seidenberg, M. S., & Binder, J. R. (2016b). Heteromodal cortical areas encode sensory-motor features of word meaning. *Journal of Neuroscience*, *36*(38), 9763−9769. https://doi.org/10.1523/jneurosci.4095-15.2016

Fernandino, L., Tong, J.-Q., Conant, L. L., Humphries, C. J., & Binder, J. R. (2022). Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(6). https://doi.org/10.1073/pnas.2108091119

Finn, E. S., & Bandettini, P. A. (2021). Movie-watching outperforms rest for functional connectivity-based prediction of behavior. *NeuroImage*, *235*, 117963. https://doi.org/10.1016/j.neuroimage.2021.117963

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimers Disease*, *49*(2), 407−422. https://doi.org/10.3233/jad-150520

Gao, Z., Zheng, L., Gouws, A., Krieger-Redwood, K., Wang, X., Varga, D., ... & Jefferies, E. (2023). Context free and context-dependent conceptual representation in the brain. *Cerebral Cortex*, *33*(1), 152−166. https://doi.org/10.1093/cercor/bhac058

Goldberg, Y. (2019). Assessing BERT's syntactic abilities. *arXiv*. https://doi.org/10.48550/arXiv.1901.05287

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369−380. https://doi.org/10.1038/s41593-022-01026-4

Gonzalez, J., Barros-Loscertales, A., Pulvermüller, F., Meseguer, V., Sanjuan, A., Belloch, V., & Avila, C. (2006). Reading cinnamon activates olfactory brain regions. *NeuroImage*, *32*(2), 906−912. https://doi.org/10.1016/j.neuroimage.2006.03.037

Graves, A., Mohamed, A.-r., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645−6649, Vancouver, BC, Canada. https://doi.org/10.1109/ICASSP.2013.6638947

Hagoort, P., & Indefrey, P. (2014). The neurobiology of language beyond single words. *Annual Review of Neuroscience*, *37*(1), 347−362. https://doi.org/10.1146/annurev-neuro-071013-013847

Hamilton, L. S., & Huth, A. G. (2020). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*(5), 573582. https://doi.org/10.1080/23273798.2018.1499946

Harris, Z. S. (1954). Distributional structure. *Word-Journal of the International Linguistic Association*, *10*(2-3), 146−162. https://doi.org/10.1080/00437956.1954.11659520

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, *28*(10), 2539−2550. https://doi.org/10.1523/JNEUROSCI.5487-07.2008

Hobbs, J. R. (1977). Pronoun resolution. *ACM SIGART Bulletin (*61), 28−28. https://doi.org/10.1145/1045283.1045292

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735−1780. https://doi.org/10.1162/neco.1997.9.8.1735

Humphreys, G. F., Lambon Ralph, M. A., & Simons, J. S. (2021). A unifying account of angular gyrus contributions to episodic and semantic cognition. *Trends in Neurosciences*, *44*(6), 452−463. https://doi.org/10.1016/j.tins.2021.01.006

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453−458. https://doi.org/10.1038/nature17637

Jain, S., & Huth, A. G. (2018, December). Incorporating context into language encoding models for fMRI. *Advances in Neural Information Processing Systems*, *31*, 6629−6638, Montreal, Canada.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169−15211. https://doi.org/10.1007/s11042-018-6894-4

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex*, *48*(7), 805−825. https://doi.org/10.1016/j.cortex.2011.04.006

Kivisaari, S. L., van Vliet, M., Hulten, A., Lindh-Knuutila, T., Faisal, A., & Salmelin, R. (2019). Reconstructing meaning from bits of information. *Nature Communications*, *10*(1), 927. https://doi.org/10.1038/s41467-019-08848-0

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148−1160. https://doi.org/10.1038/s41593-018-0210-5

Kriegeskorte, N., & Douglas, P. K. (2019). Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, *55*, 167−179. https://doi.org/10.1016/j.conb.2019.04.002

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, *28*(1), 40−80. https://doi.org/10.3758/s13423-020-01792-x

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23−49. https://doi.org/10.1016/j.brainres.2006.12.063

Lahat, D., Adali, T., & Jutten, C. (2015). Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, *103*(9), 1449−1477, . https://doi.org/10.1109/jproc.2015.2460697

Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews: Neuroscience*, *18*(1), 42−55. https://doi.org/10.1038/nrn.2016.150

Laurino Dos Santos, H., & Berger, J. (2022). The speed of stories: Semantic progression and narrative success. *Journal of Experimental Psychology: General*, *151*(8), 1833−1842. https://doi.org/10.1037/xge0001171

Law, R., & Pylkkanen, L. (2021). Lists with and without syntax: A new approach to measuring the neural processing of syntax. *Journal of Neuroscience*, *41*(10), 2186−2196. https://doi.org/10.1523/JNEUROSCI.1179-20.2021

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788−791. https://doi.org/10.1038/44565

Lee, H., & Chen, J. (2022). Predicting memory from the network structure of naturalistic events. *Nature Communications*, *13*(1), 4235. https://doi.org/10.1038/s41467-022-31965-2

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*(8), 2906−2915. https://doi.org/10.1523/jneurosci.3684-10.2011

Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., ... Smallwood, J. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(44), 12574−12579. https://doi.org/10.1073/pnas.1608282113

Matchin, W., Brodbeck, C., Hammerly, C., & Lau, E. (2019). The temporal dynamics of structure and content in sentence comprehension: Evidence from fMRI-constrained MEG. *Human Brain Mapping*, *40*(2), 663−678. https://doi.org/10.1002/hbm.24403

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schutze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(42), 25966−25974. https://doi.org/10.1073/pnas.1910416117

Mcculloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115−133. https://doi.org/10.1007/bf02478259

Miani, A., Hills, T., & Bangerter, A. (2022). Interconnectedness and (in)coherence as a signature of conspiracy worldviews. *Science Advances*, *8*(43), eabq3668. https://doi.org/10.1126/sciadv.abq3668

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv*. https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Karafiat, M., Burget, L., Cernocky, J. H., & Khudanpur, S. (2010, September). Recurrent neural network based language model. *11th Annual Conference of the International Speech Communication Association 2010*, 1045−1048, Makuhari, Chiba, Japan.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *arXiv*. https://doi.org/10.48550/arXiv.1310.4546

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191−1195. https://doi.org/10.1126/science.1152876

Nastase, S. A., Connolly, A. C., Oosterhof, N. N., Halchenko, Y. O., Guntupalli, J. S., Visconti Di Oleggio Castello, M., ... Haxby, J. V. (2017). Attention selectively reshapes the geometry of distributed semantic representation. *Cerebral Cortex*, *27*(8), 4277−4291. https://doi.org/10.1093/cercor/bhx138

Nevler, N., Ash, S., McMillan, C., Elman, L., McCluskey, L., Irwin, D. J., ... Grossman, M. (2020). Automated analysis of natural speech in amyotrophic lateral sclerosis spectrum disorders. *Neurology*, *95*(12), E1629−E1639. https://doi.org/10.1212/wnl.0000000000010366

Nguyen, M., Vanderwal, T., & Hasson, U. (2019). Shared understanding of narratives is correlated with shared neural responses. *NeuroImage*, *184*, 161−170. https://doi.org/10.1016/j.neuroimage.2018.09.010

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology / Revue canadienne de psychologie*, *45*(3), 255−287. https://doi.org/10.1037/h0084295

Patel, T., Morales, M., Pickering, M. J., & Hoffman, P. (2022). A common neural code for meaning in discourse production and comprehension. *bioRxiv*. https://doi.org/10.1101/2022.10.15.512349

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976−987. https://doi.org/10.1038/nrn2277

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018, June). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 2227−2237, New Orleans, Louisiana, USA. https://doi.org/10.18653/v1/N18-1202

Petersson, K.-M., Folia, V., & Hagoort, P. (2012). What artificial grammar learning reveals about the neurobiology of syntax. *Brain and Language*, *120*(2), 83−95. https://doi.org/10.1016/j.bandl.2010.08.003

Poldrack, R. A., Wagner, A. D., Prull, M. W., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. (1999). Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. *NeuroImage*, *10*(1), 15−35. https://doi.org/10.1006/nimg.1999.0441

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022). Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife*, *11*, e77599. https://doi.org/10.7554/elife.77599

Pulvermüller, F. (2013). How neurons make meaning: Brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, *17*(9), 458−470. https://doi.org/10.1016/j.tics.2013.06.004

Pulvermüller, F., Harle, M., & Hummel, F. (2001). Walking or talking? Behavioral and neurophysiological correlates of action verb processing. *Brain and Language*, *78*(2), 143−168. https://doi.org/10.1006/brln.2000.2390

Pulvermüller, F., Kherif, F., Hauk, O., Mohr, B., & Nimmo-Smith, I. (2009). Distributed cell assemblies for general lexical and category-specific semantic processing as revealed by fMRI cluster analysis. *Human Brain Mapping*, *30*(12), 3837−3850. https://doi.org/10.1002/hbm.20811

Pulvermüller, F., Lutzenberger, W., & Preissl, H. (1999). Nouns and verbs in the intact brain: Evidence from event-related potentials and high-frequency cortical responses. *Cerebral Cortex*, *9*(5), 497−506. https://doi.org/10.1093/cercor/9.5.497

Pylkkanen, L. (2019). The neural basis of combinatory syntax and semantics. *Science*, *366*(6461), 62−66. https://doi.org/10.1126/science.aax0050

Quoc, L., & Mikolov, T. (2014, June). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning*, *32*, 1188−1196, Beijing, China.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *46*(2), 245−266. https://doi.org/10.1006/jmla.2001.2810

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613−620. https://doi.org/10.1145/361219.361220

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(45). https://doi.org/10.1073/pnas.2105646118

Schwartz, D., Toneva, M., & Wehbe, L. (2019, December). Inducing brain-relevant bias in natural language processing models. *Advances in Neural Information Processing Systems*, *32*, 14123−14133, Vancouver, Canada. https://dl.acm.org/doi/10.5555/3454287.3455553

Segaert, K., Menenti, L., Weber, K., Petersson, K. M., & Hagoort, P. (2012). Shared syntax in language production and language comprehension—an fMRI study. *Cerebral Cortex*, *22*(7), 1662−1670. https://doi.org/10.1093/cercor/bhr249

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, *7*(1), 12141. https://doi.org/10.1038/ncomms12141

Smallwood, J., Bernhardt, B. C., Leech, R., Bzdok, D., Jefferies, E., & Margulies, D. S. (2021). The default mode network in cognition: A topographical perspective. *Nature Reviews Neuroscience*, *22*(8), 503−513. https://doi.org/10.1038/s41583-021-00474-4

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631−1642, Seattle, Washington, USA. https://aclanthology.org/D13-1170

Solomon, S. H., Medaglia, J. D., & Thompson-Schill, S. L. (2019). Implementing a concept network model. *Behavior Research Methods*, *51*(4), 1717−1736. https://doi.org/10.3758/s13428-019-01217-1

Sun, X., Yang, D., Li, X., Zhang, T., Meng, Y., Qiu, H., ... Li, J. (2021). Interpreting deep learning models in natural language processing: A review. *arXiv*. https://doi.org/10.48550/arXiv.2110.10470

Sundermeyer, M., Schluter, R., & Ney, H. (2012, September). LSTM neural networks for language modeling. *13th Annual Conference of the International Speech Communication Association*, 194−197, Portland, Oregon, USA. https://doi.org/10.21437/Interspeech.2012-65

Sutskever, I., Vinyals, O., & Le, Q. V. (2014, December). Sequence to sequence learning with neural networks. *Advances*

in Neural Information Processing Systems, *27*, Montreal, Canada. https://dl.acm.org/doi/10.5555/2969033.2969173

Tikochinski, R., Goldstein, A., Yeshurun, Y., Hasson, U., & Reichart, R. (2021). Fine-tuning of deep language models as a computational framework of modeling listeners' perspective during language comprehension. *bioRxiv*. https://doi.org/10.1101/2021.11.22.469596

Toneva, M., & Wehbe, L. (2019, December). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, *32*, 14954−14964, Vancouver, Canada. https://dl.acm.org/doi/10.5555/3454287.3455626

Tong, J., Binder, J. R., Humphries, C., Mazurchuk, S., Conant, L. L., & Fernandino, L. (2022). A distributed network for multimodal experiential representation of concepts. *Journal of Neuroscience*, *42*(37), 7121−7130. https://doi.org/10.1523/JNEUROSCI.1243-21.2022

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017, December). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, Long Beach, California, USA. https://dl.acm.org/doi/10.5555/3295222.3295349

Vodrahalli, K., Chen, P.-H., Liang, Y., Baldassano, C., Chen, J., Yong, E., ... Arora, S. (2018). Mapping between fMRI responses to movies and their natural language annotations. *NeuroImage*, *180*, 223−231. https://doi.org/10.1016/j.neuroimage.2017.06.042

Wang, S., Zhang, J., Lin, N., & Zong, C. (2018, February). Investigating inner properties of multimodal representation and semantic compositionality with brain-based componential semantics. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1), 5964−5972, New Orleans, Louisiana, USA. https://doi.org/10.1609/aaai.v32i1.12032

Wang, S., Zhang, J., Lin, N., & Zong, C. (2020, February). Probing brain activation patterns by dissociating semantics and syntax in sentences. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 9201−9208, New York, USA. https://doi.org/10.1609/aaai.v34i05.6457

Wang, S., Zhang, J., & Zong, C. (2018, October-November). Associative multichannel autoencoder for multimodal word representation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 115−124, Brussels, Belgium. https://doi.org/10.18653/v1/D18-1011

Wang, X., Wu, W., Ling, Z., Xu, Y., Fang, Y., Wang, X., ... Bi, Y. (2018). Organizational principles of abstract words in the human brain. *Cerebral Cortex*, *28*(12), 4305-4318. https://doi.org/10.1093/cercor/bhx283

Warburton, E., Wise, R. J., Price, C. J., Weiller, C., Hadar, U., Ramsay, S., & Frackowiak, R. S. (1996). Noun and verb retrieval by normal subjects studies with PET. *Brain*, *119*, 159−179. https://doi.org/10.1093/brain/119.1.159

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS One*, *9*(11), e112575. https://doi.org/10.1371/journal.pone.0112575

Wu, S., & Dredze, M. (2019, November). Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 833−844, Hong Kong, China. https://doi.org/10.18653/v1/D19-1077

Wurm, M. F., & Caramazza, A. (2019). Distinct roles of temporal and frontoparietal cortex in representing actions across vision and language. *Nature Communications*, *10*(1), 289. https://doi.org/10.1038/s41467-018-08084-y

Xu, C., Zhang, Y., Zhu, G., Rui, Y., Lu, H., & Huang, Q. (2008). Using webcast text for semantic event detection in broadcast sports video. *IEEE Transactions on Multimedia*, *10*(7), 1342−1355. https://doi.org/10.1109/Tmm.2008.2004912

Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic Bulletin & Review*, *23*(4), 1015−1027. https://doi.org/10.3758/s13423-015-0948-7

Yeshurun, Y., Nguyen, M., & Hasson, U. (2021). The default mode network: Where the idiosyncratic self meets the shared social world. *Nature Reviews: Neuroscience*, *22*(3), 181−192. https://doi.org/10.1038/s41583-020-00420-w

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of cnn and rnn for natural language processing. *arXiv*. https://doi.org/10.48550/arXiv.1702.01923

Zhang, X., Wang, S., Lin, N., Zhang, J., & Zong, C. (2022, February). Probing word syntactic representations in the brain by a feature elimination method. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(10), 11721−11729. https://doi.org/10.1609/aaai.v36i10.21427

Zhang, Y., & Wallace, B. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv*, 253−263. https://doi.org/10.48550/arXiv.1510.03820

Zhu, X., Li, Z., Wang, X., Jiang, X., Sun, P., Wang, X., ... Yuan, N. J. (2022). Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 1−20. https://doi.org/10.1109/tkde.2022.3224228

Distributed representation of semantics in the human brain: Evidence from studies using natural language processing techniques

JIANG Jiahao[1], ZHAO Guoyu[2], MA Yingbo[1], DING Guosheng[3], LIU Lanfang[2,4]

([1] *Faculty of Psychology, Beijing Normal University, Beijing* 100875, *China*)

([2] *Faculty of Psychology, School of Arts and Sciences, Beijing Normal University, Zhuhai* 519087, *China*)

([3] *State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University & IDG/McGovern Institute for Brain Research, Beijing* 100875, *China*) ([4] *Center for Cognition and Neuroergonomics at the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Zhuhai* 519087, *China*)

Abstract: How semantics are represented in human brain is a central issue in cognitive neuroscience. Previous studies typically address this issue by artificially manipulating the properties of stimuli or task demands. Having brought valuable insights into the neurobiology of language, this psychological experimental approach may still fail to characterize semantic information with high resolution, and have difficulty quantifying context information and high-level concepts. The recently-developed natural language processing (NLP) techniques provide tools to represent the discrete semantics in the form of vectors, enabling automatic extraction of word semantics and even the information of context and syntax. Recent studies have applied NLP techniques to model the semantic of stimuli, and mapped the semantic vectors onto brain activities through representational similarity analyses or linear regression. A consistent finding is that the semantic information is represented by a vastly distributed network across the frontal, temporal and occipital cortices. Future studies may adopt multi-modal neural networks and knowledge graphs to extract richer information of semantics, apply NLP models to automatically assess the language ability of special groups, and improve the interpretability of deep neural network models with neurocognitive findings.

Keywords: semantic representation, brain, natural language processing, language model